Game-based Assessment of Psychoacoustic Thresholds: Not All Games Are Equal!

Vero Vanden Abeele[a], Jan Wouters[b], Pol Ghesquière[c], Ann Goeleven[d] and Luc Geurts[a]
[a] e-Media Lab, KU Leuven, Leuven, Belgium
{vero.vandenabeele; luc.geurts}@kuleuven.be
[b] ExpORL, Department of Neurosciences, KU Leuven, Leuven, Belgium
jan.wouters@med.kuleuven.be
[c] Parenting and Special Education Research Unit, KU Leuven, Leuven, Belgium
pol.ghesquiere@ppw.kuleuven.be
[d] Dept. Speech Language Pathology (MUCLA), University Hospitals Leuven, Belgium
ann.goeleven@uzleuven.be

ABSTRACT
This paper first presents a critical analysis of an existing game (APEX), designed by researchers in psychoacoustics only, to measure psychoacoustic thresholds in preschoolers. Next it presents another game (DIESEL-X), designed by dyslexia researchers *and* game designers, to remediate the shortcomings of the former game. Via a repeated measures experiment (n = 95), the game experience, attention, and psychoacoustic thresholds are compared. It is shown that the children prefer the game experience of DIESEL-X over APEX. Moreover, the former game was able to measure lower frequency-modulation thresholds than APEX. These results demonstrate that when it comes down to game-based assessment of children's perceptual capabilities, the quality of game design not only has an effect on game experience, but equally on the scientific measurements obtained via such a game-based assessment.

Author Keywords
Dyslexia; game-based assessment; psychometrics; serious games; psychoacoustics; staircase method; perceptual thresholds.

ACM Classification Keywords
H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

INTRODUCTION

Psychoacoustics and Games
Psychoacoustics is a sub-discipline of psychophysics where the relation between acoustic stimuli and subjective responses are investigated. Experiments typically require participants to listen to one or more sounds and to make a judgment based on a certain perceptual quality of those sounds, such as loudness, pitch or any other alteration of sounds. In order to come to valid conclusions, many judgments have to be made, resulting in lengthy tests that might be perceived as tiresome or boring. In case the participants are young children, the challenge for the researcher supervising the experiments accrues. In order to hold the attention of the child throughout the tests, the experimenter needs to take special actions to motivate the child: praising continuously, giving extra rewards such as sweets and stickers, telling jokes, etc.

In this paper we investigate whether games can offer a reliable method to perform psychoacoustic measurements with young children. More specifically, we discuss two games that were designed to measure the threshold of frequency-modulated tones in preschoolers (5 years old). One game (APEX) was designed by psychoacoustic researchers, i.e. non-experts in game design. The other game (DIESEL-X) was designed and developed by a multidisciplinary team including domain experts *and* game experts. Three aspects were investigated and compared: the children's subjective preference for one of the games, the attention span for both games and the psychoacoustic thresholds. Based on these psychoacoustic thresholds, the games aimed to find measures to predict whether a preschooler has a high risk for developing dyslexia.

In the next section we explain the link between dyslexia and basic psychoacoustic measurements. Afterwards, more detailed information on the specific procedures of the psychoacoustic experiment is given. Then, both the APEX game and DIESEL-X game are presented and analyzed according to principles of 'good' game design. Next, we present an experimental evaluation of both games with 95 preschoolers, where the game experience, attention, and psychoacoustic thresholds are compared. These results demonstrate that when it comes down to game-based assessment of children's perceptual capabilities, not all games are equal. Finally, these results are discussed and future work is suggested.

CHI PLAY 2015, October 03 - 07, 2015, London, United Kingdom
 ISBN 978-1-4503-3466-2/15/10…$15.00
DOI: http://dx.doi.org/10.1145/2793107.2793132

Dyslexia and Psychoacoustics

Dyslexia is a developmental reading disorder that is defined by *"persistent difficulties in reading and writing fluently, despite normal or above-average intelligence and despite active remediation"* [19,43]. As with many developmental disorders, it should be detected as soon as possible. The younger the age at the start of a treatment, the larger the effect that can be attained [17,22,27,30]. Hence, ideally, the risk for dyslexia is assessed *before* formal reading and writing instruction take place.

In the past decade, there has been a growing consensus among dyslexia researchers that a phonological deficit is causal to these specific reading and spelling errors. The *phonological deficit theory* postulates that problems originate from a deficit that is specific to the phonological representation and processing of speech sounds [1,45]. Children with dyslexia seem to be less sensitive for the sound structure of language – which is for example needed to recognize rhyming words, or words starting or ending with the same sound [10]. In turn this phonological deficit is assumed to be caused by underlying neurological dysfunctioning, in particular by difficulties in low-level auditory temporal processing. Children with dyslexia tend to have difficulties processing linguistic *and* nonlinguistic stimuli that are short and enter the nervous system in rapid succession [1,16,32]. Children with dyslexia show an impaired perception of dynamic aspects in the auditory signal itself, like amplitude and frequency modulations [34,46,47,50]. This implies that psychoacoustic tests, that do not require reading or writing, do allow for the detection of high risk for dyslexia, even at preschool age [5–9].

In a series of experiments, Boets and colleagues assessed phonological ability, speech perception and low-level auditory processing in both a group of 5-year-old pre-school children at high risk for dyslexia[1], compared to a group of well-matched 5-year-old children at low risk for dyslexia [5–9]. The performance on these tests yielded a good predictor for the development of dyslexia; the high risk dyslexia group scored significantly worse on the test than the low risk dyslexia group. However, it was not possible to make predictions at the individual level. The test results showed a lot of variability for threshold values where the 'best performance' of the child was to be measured. The involved researchers remarked that although these experiments were incorporated in a game (APEX, detailed further in the paper) [6,25], it was still difficult to keep the child's attention at a high level throughout the test.

The researchers concluded that better measurements were needed to allow for risk detection at the individual level. Consequently, a new project was devised that aimed to design and develop a game for preschoolers, that allows to predict at an individual level whether the child has a high risk for developing dyslexia. The resulting game is DIESEL-X. It consists of three mini-games that are described in [12,20]. One mini-game specifically intends to measure the threshold for frequency-modulation detection and is the focus of this paper.

Adaptive procedures in Psychoacoustics

Typically, threshold measurements in psychoacoustics are done via adaptive procedures. An adaptive procedure is one in which the stimulus level in one trial is determined by the preceding stimuli and responses [28]. Most typical is the method of up and down, or the staircase-method (Figure 1), as the curve takes the shape of a staircase. More specifically, a 2-up, 1-down method is used in our experiments, meaning that the difficulty level is increased after two correct responses, and decreased after one incorrect response. The task is to pick out the odd stimulus in a series of three consecutive tones. Two tones are unmodulated, pure sinusoidal tones, and one tone is frequency modulated, with the depth varying from trial to trial. At the start of the experiment, this depth is very large and thus easily discernable. Figure 1 illustrates a typical evolution of the "difficulty level" (vertical axis) when the experiment progresses (trial number on horizontal axis). Around the threshold value, so called "reversals" take place, when a series of correct responses is followed by an incorrect response (negative reversal, indicated with a minus sign), or vice versa (positive reversal, indicated with a plus sign). The average value of the stimulus levels at the reversal is then taken as the estimate of the threshold value. In case of a 2-up 1-down method, this threshold level corresponds to a stimulus that can be detected correctly in 70.7% of the cases [28].

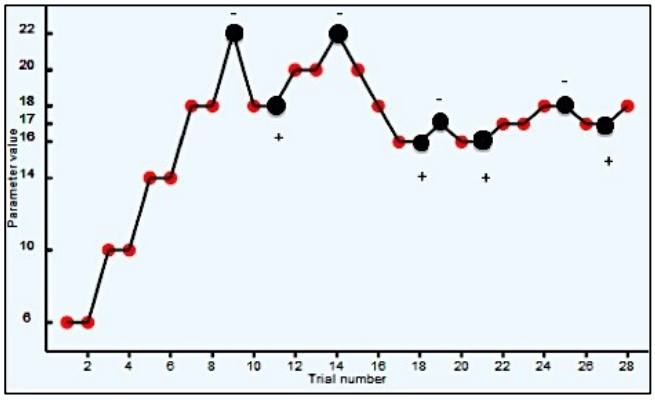

Figure 1 Example evolution of the stimulus levels throughout an adaptive experiment. A higher value on the vertical axis corresponds to a stimulus that is more difficult to detect. Reversals are indicated with plus and minus signs.

[1] Dyslexia is a hereditary disorder. Between twenty three and fifty six percent of children with dyslexia have a parent with the disorder. Hence, high-risk groups consist of children of parents with dyslexia, whereas low-risk groups consist of children with parents without dyslexia.

TWO GAMES FOR PSYCHOACOUSTIC TESTING

APEX: the first attempt

The aforementioned tests of Boets and colleagues were conducted with APEX (Application for PsychoElectrical eXperiments), a software program for psychoacoustic experiments. While in origin APEX was not specifically designed for testing children, for these experiments the classical test software had been gamified to allow testing of very young children [7,25]. It was hypothesized that better measurements can be achieved by finding better ways to engage and motivate the child to take part in the test and consequently to better grab attention [37] and hold attention [15,18,40,41]. APEX is considered to be an interactive video game by the dyslexia researchers, in that 1) APEX contains sets of intro movies and animated cartoons, 2) during tests every interval is visually represented on the screen by a funny character, that is animated synchronously with the presentation of the corresponding sound interval, 3) a correct response is reinforced with a spectacular movement of the selected character, 4) APEX rewards children by adding smiley faces to a rising ladder structure for every correct response. Every incorrect response is punished by removing a smiley from the ladder structure.

More concrete, APEX starts with a brief animation of a mother dragon looking at three eggs. In one of the eggs the mother's baby dragon is hidden. That egg sounds a bit different (modulated tone) from the two others (pure tones). One of the eggs shakes when the corresponding tone is played, so the preschooler knows which tone corresponds to which egg. When the right egg is selected, the baby dragon appears. In the other case, an empty egg is shown. Upon correctly identifying the baby dragon, players receive a smiley. When pointing to an incorrect egg, a smiley is removed.

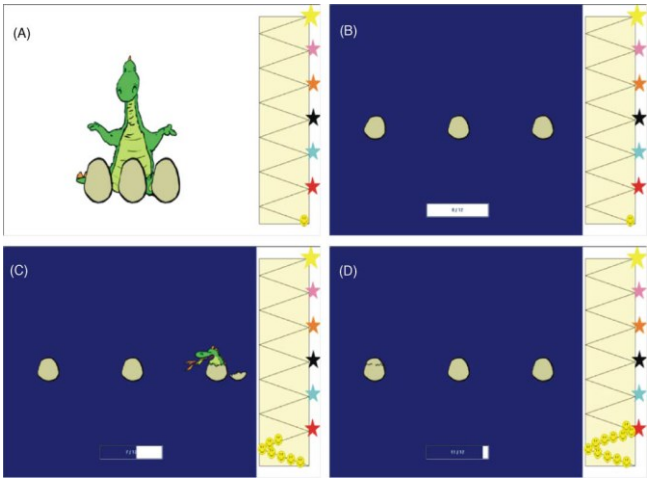

Figure 2 Four screenshots from the APEX game. a) Mother dragon is looking for her children. b) one of the three eggs contains the baby dragon, the other are empty, c) the child has correctly identified the baby dragon, d) the egg simply shows a crack, the child selected the wrong egg.

The stars (after every ten smileys) on the ladders have no consequence, they are simply used as a graphical marker. To round off the experiment and as a final reward for the child's achievement, an animated closing movie is shown.

Is APEX a (good) game?

APEX was created by dyslexia researchers without formal training in game design. It has been argued before that all too often 'serious games' lack depth and are essentially sugar coating or 'chocolate covered broccoli' [11]. Hence, it can be interesting to analyze APEX through the lens of (instructional) game design, and through the lens of intrinsic motivation [21].

Endogenous vs exogenous fantasy

Most games incorporate make-belief and evoke mental imagery. Embedding content in fantasy contexts leads to greater player interest and increased learning [13,36]. As aforementioned, APEX presents the fantasy of a mother dragon looking for her child. The narrative is introduced via the illustrations (see Figure 2) and via the following sentences: *"Mama dragon is sitting on the lawn. She is wondering where her little dragons are, because she only sees three eggs. Can you tell which egg sounds different?"*

In this context, the notion of intrinsic fantasy vs. extrinsic fantasy moves to the fore, also termed endogenous vs exogenous fantasy [4,18]. An exogenous or extrinsic fantasy is simply overlaid on the game content. An endogenous or intrinsic fantasy is related to the skill required of the player. An example of an extrinsic fantasy is doing tables of multiplication, in order to be able to shoot at a zombie. There is no link between the skill of being good in the tables of multiplication, and being able to shoot zombies. An example of an intrinsic fantasy is a detective that needs to decipher a numeric code by doing performing calculations. In this last example, there is an essential relationship between the learned skill and the fantasy context; there is a link between deciphering a numeric code and the skill of performing multiplications. In the APEX game, the fantasy provides no direct link between the skill (detecting frequency modulation) and the discovery of the baby dragon.

Challenges, adaptivity and perceived competence

Setting challenges at the right level is important for perceived competence and the self-esteem of a player. Success in a game makes players feel better of themselves. Self-esteem is essential to motivation to perform a behavior [2,15]. The opposite principle of this implies that failure in a game may lower a person's self-esteem. Because of the 2-up, 1-down staircase procedure, the APEX game is bound to progress to and oscillate around the threshold level at which the child can detect the correct stimulus, at a chance level of 70.7% [28]. Such an adaptive and progressive difficulty that typifies the staircase method aligns with the core concept of flow; for an optimal game experience, the challenges should match the skills of the player [14]. This is also reflected in the notion of hard fun [26] or Bushnell's theorem that a game should be easy to learn but hard to master [51].

However, note that with psychometric testing, there is little that can be learned or 'mastered' by the player over the course of a game. In fact, psychometric tests are based on the assumption that *"the psychometric functions are stationary with time,"* [28], i.e., there can be no change in the functioning of the player during the course of a test. Hence, if learning or mastery would occur, this would violate the above assumption and render the psychometric test invalid. This points to a crucial difference between games for assessment and games for learning/training, and unveils a fundamental problem that all games that aim at assessment have to deal with. The APEX game as all *assessment games is measuring a skill that is (or should be) beyond the control of the player.*

Hence, while APEX is adaptive, one can question the amount of learner control, i.e. the capacity of a player to influence or exert power over certain elements of the game [18]. Learner control is essential to a sense of mastery or perceived competence but also for a sense of autonomy, both ingredients of self-determination and intrinsic motivation [15].

Rewards, self-efficacy and autonomy

Moreover, APEX has a direct one-on-one mapping between the positive response (the correct identification of the baby dragon) and the reward/punishment. APEX provides children a smiley when having successfully identified the egg with the baby dragon. When pointing at the wrong egg, the smiley is removed. As the child success rate converges on 70.7%, in roughly one out of three attempts, a smiley is removed. In any game there is a thin line between "*the need to provide clear performance feedback to enhance the challenge, and the need to not reduce self-esteem to the point where the challenge becomes discouraging rather than inviting*"[31]. While there are no clear guidelines on where that line should be (i.e. what the success rate should be), the removal of smileys should be questioned, because of the negativity bias [42]. People, and children in particular, are more sensitive to punishment than rewards [3], and this may further undermine feelings of self-efficacy [2] and control or autonomy [15].

Goals, performance feedback and meaningful play

Essential for any game is a clear and attainable goal [29]. Clear, specific goals allow the player to perceive goal-feedback discrepancies, which are crucial in triggering greater attention and motivation [18]. In APEX the goal is identifying the baby dragon, with a direct visual feedback of the baby dragon appearing out of the egg. However, not all goals are equal. Important is to structure multiple levels of goals, and to offer a meta-goal [31]. Salen & Zimmerman (2004) talk about the importance of discernibility and integrated play for meaning play: *"Meaningful play occurs when the relation between actions and outcomes are discernable and integrated into the larger context of the game."* Whereas *discernibility* emphasizes the relation between the small actions and the feedback on every little action (pointing at the egg, and the resulting appearance of the baby dragon), *integration* emphasizes that it should be clear how every action contributes to the larger goal of the game. Such a meta-goal is not implemented in APEX and may influence the overall significance that the player attributes to the game.

Curiosity, interest and exploration

Curiosity is a final characteristic that is stressed when designing for intrinsic motivation [35] and game experiences [48]; games should be novel and surprising. Garris et al. talk about the need for mystery as the gap between existing information and unknown information [18]. Malone distinguishes between cognitive curiosity (the desire to bring better form to one's knowledge structures), and sensory curiosity implies to the use of audio and visual effects to enhance feelings of novelty and exploration. Sensory curiosity in the shape of audio and visual effects increases the sense of fantasy, and increase the salience of performance feedback (i.e. it strengthens the reward). APEX is limited in curiosity and mystery. After the initial introductory movie, every trial is identical to the one before, the animations and sounds are identical for every trial, when being reward the same smiley is given. After every ten trials, the smileys on the ladder reach a star, but no extra animation is given. No extra rewards can be gained, no extra levels unlocked. Only after eight reversals, the game suddenly ends and shows a final animation.

Conclusion

APEX was designed as an interactive video game to find better ways to engage and motivate the child to take part in the psychoacoustic test and consequently to attain a longer attention span. However, researchers, when using APEX, noted that preschoolers still lacked motivation and had trouble maintaining attention throughout the test. Upon analyzing APEX through the lens of game design and intrinsic motivation, we noted that has its shortcomings with respect to meaningful play (lack of a meta-goal), learner control (lack of control over the outcome), endogenous fantasy (lack of an essential relationship between skill and in game action) and mystery (a lack of cognitive and sensory curiosity). Perhaps, the aforementioned shortcomings explain why in testing procedures, much responsibility was still given to the researcher, who prompted the child to be ready, verbally and by pointing at the screen. In addition, researchers lauded the child for its effort, and provided real life stickers in addition to the smileys on the ladder structure.

DIESEL-X: a better game?

The APEX researchers turned to our research group with the question to design a 'better' game: a game that motivates preschoolers and hold their attention for a longer time, and as a result can more reliably measure psychoacoustic thresholds. In addition, the researchers also hoped that this game would no longer require the active presence of a researcher. Ideally, such a game could be played from start to end, without any intervention of an adult.

Figure 3. The main character of the game is Diesel, a police robot dog that helps Alex, a tomboy, to chase nasty cats that are causing havoc in the town.

Over the course of 18 months (October 2011 – March 2013) and via a player-centered design process [49], we developed such an 'enhanced' computer game. We observed and interviewed preschoolers with respect to their most preferred games [20] by means of a diary study. We conducted a laddering study to unveil the gameplay preferences of preschoolers [12] on reward structures, input mechanisms, character creation... We involved preschoolers in the selection of the theme and narrative via focus groups. We tested game prototypes iteratively and incrementally during the development process. The full player-centered design and development of the game is beyond the scope of this paper but is discussed at length in [12,20]. In the following paragraphs, we provide the details of the game as deemed relevant for this paper.

The DIESEL-X game starts with an animated story of a gang of bad cats stealing money and diamonds from the bank. The main character Alex is introduced, who has the help of a robot police dog, called Diesel (Fig. 3). An intro animation shows how the cats rob a bank, and how the local town sheriff expresses his helplessness. The child's task is to help Diesel to retrieve all the money and the three diamonds, by

Figure 4. Screen shot of the FM detection task. Diesel is chasing three cats carrying bags. One of the bags contains stolen money or diamonds. When the trial starts, a sound is played for each cat. An FM tone is played for the 'bad' cat (correct response), and a pure tone is played for the other two cats (incorrect response).

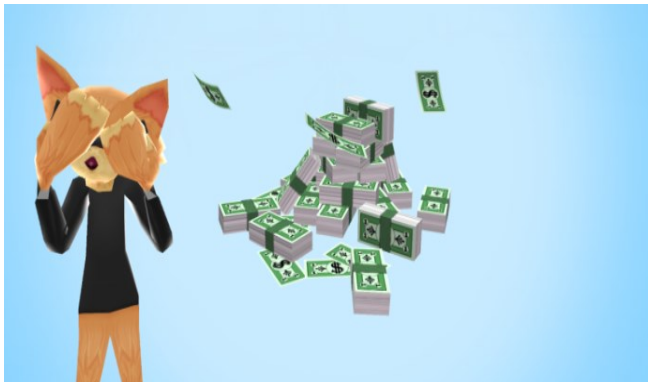

Figure 5. When the child identifies the correct bag, the cat has to return the money, much to his dismay.

catching the cats. Luckily, being a robot police dog, Diesel has a police scanner. Before each trial, Diesel is chasing three cats through the city streets (Fig. 4): One cat carries a bag with money or diamonds (modulated tone), the two others are carrying stones (pure tones) in order to mislead Diesel. Diesel has a police scanner that when aimed at the cats can makes a specific sound in case the bag contains money. After the sounds are played, the child answers by pointing at the cat on the tablet's touch screen. When catching a cat, there is a moment of surprise where the child is not sure whether the bag will contain bricks (incorrect choice) or money (correct response). When the child identifies the correct bag, the cat has to return the money, to his dismay (Fig. 5). After two reversals, the child is rewarded in an extra manner. The bag not only contains money but a diamond as well. Upon having caught a diamond, the child is now able to give Diesel a new color, or to select extra gear for Diesel, such as a jet pack or propellers (Fig. 6). The game ends when all the money has been retrieved and the cats are caught. They are then send off to the moon in a rocket, never to return (Fig. 7). Piece and calm return in the city, and Alex and Diesel are praised by the sheriff for their courage.

Analysis of DIESEL-X

Goals, performance feedback and meaningful play

The team chose the theme and meta-goal in collaboration with preschoolers, and this meta-goal is clearly integrated in

Figure 6. After two reversals, the child is allowed to choose an extra color for Diesel, or to select extra gear such as a jet pack or propellers.

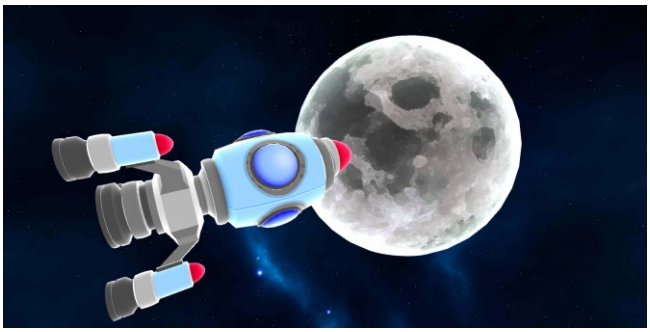

Figure 7 The cats are sent off to the moon, never to return.

the overall game. The meta-goal is also introduced in the beginning of the game, namely chasing cats, and restoring piece in the city. Upon every action (identifying the bag with the stolen money), the content of the bag is shown (bricks or money/diamonds) and the reaction of the cat. Moreover, after run 2, 4, and 6 (because of the structure of the game every second reversal is triggered when the child performs two consecutive good actions) the child receives not only money but a diamond as well. After finding a diamond the child unlocks a new color or gear for Diesel.

When three diamonds are found, the game is over (all the stolen money and diamonds are found). Hence the game is comprised of multiple structured goals. The goals vary from retrieving money, to retrieving diamonds and finally sending the cats off in space. There are no punishments, no money or diamonds are taken away from the child. In this way we aim to foster attachment of the player to the outcome of the game, in other words meaningful play.

Challenges, adaptivity and learner control

As aforementioned, the scientific procedure of the staircase method does not allow to tinker with difficulties nor the chance level of positive responses. Moreover, the game measures a psychometric function, hence no real learning or mastery can take place. This poses a challenge with respect to mastery. To provide a way to increase learner control and autonomy, we added an extra scene where the child is rewarded for having collected a diamond by unlocking an extra color or extra gear for Diesel. The child can try out this new color or gear, but can choose equally from other colors/gear. While this has no influence on the actual performance on the tests, this does give the child freedom of choice, and perhaps a feeling of control and autonomy. In the game scenes after the child has made his choice, Diesel is wearing this newly selected color or gear. Hence, the child is given power over Diesel.

Endogenous vs exogenous fantasy

The DIESEL-X game aimed at offering a more immersive game world. The narrative introduces both Alex and Diesel at length, in the setting of the town. The narrative also shows the sheriff asking for help. The aim is that the child can more easily identify with Alex and Diesel. Moreover, by linking the skill of detecting frequency modulation to the skill of using a police scanner, we aimed for endogenous fantasy.

Curiosity and exploration

Finally, the game fosters sensory curiosity by funny responses of cats when caught, and offers clear visualizations of rewards. The extra level where children can choose novel colors and/or gears not only aims to increase a feeling of control, but equally of curiosity, as the child does not know what the new color will look like or which gear will become unlocked.

Conclusion

When analyzing DIESEL-X through the lens of game design, we aimed at meaningful play by adding a narrative and a layered goal-structure, we aimed at learner control by giving the child control over the creation of the main character. We equally aimed at providing mystery and surprise by letting the child unlock extra gear and colors for the main character. Finally, we aimed at providing an endogenous fantasy by giving the robot-police dog a police scanner, necessary for the FM modulation.

Research questions

Our hypothesis was that DIESEL-X would be able to remediate the shortcomings of the earlier game. In particular, we had the following hypotheses:

- H1. Preschoolers will have a 'better' game experience' when playing the DIESEL-X game than when playing the APEX game.
- H2. Preschoolers will show greater attention in the DIESEL-X game than the APEX game.
- H3. The DIESEL-X game will measure lower FM-thresholds than the APEX game.

To test these hypotheses, we compared the newly developed DIESEL-X game to the APEX game through an intra-subject analysis in a preschool population. Both the preschoolers' preferences and attention were investigated, as well as their performance on an FM detection task.

METHODS AND MATERIALS

Participants

95 children from the final year in preschool participated in this study (female: 43, male: 52, average age: 5 years and 10 months). The APEX-software [25] was run on a laptop, the DIESEL-X game was played on a Samsung tablet. A calibrated headset was used to play the sounds, to ensure a proper and controlled sound level on both games.

Experimental design and procedure

Upon entering the test room (situated at the school premise) and before the actual test took place, first the right ear's sensitivity to sounds was measured. In both games the sounds were presented to this ear. Measuring the ear's sensitivity is necessary to prevent that judgmental errors could be due to hearing loss, leading to an incorrect interpretation of the test results. The hearing threshold was measured at 500 Hz, 1000 Hz, 2000 Hz and 4000 Hz

according to the Hughson-Westlake method. The criterion for participation was that hearing loss did not exceed 30 dB HL. This criterion was met for all participants.

A repeated measures design was carried out with two conditions, playing the APEX game and playing the DIESEL-X game. The order of conditions was counterbalanced to rule out order effects. After having played both games, the child was asked for his preferred game. One experimental session lasted between 45 and 60 minutes per child.

Measurements

Measurement of game experience.

It would have been ideal to measure game experience [23] or intrinsic motivation [33] as multi-dimensional constructs, and by means of existing, validated surveys. However, these surveys are designed for adults and demand that test participants can read. Adapted scales for children do exit, such as the smiley-o-meter [38] and the fun-toolkit [39], but they have shown to become unreliable when used at ages of younger than eight [44,54]. At young ages, it is advisable to treat a game experience as a uni-dimensional construct [55], and to use preference evaluation methods, i.e. asking the child to compare two conditions and to indicate which one is preferred [52]. Hence, for the evaluation of the child's game experience, the This-or-That method was used, since this method has been validated with preschoolers [55].

Table 1. The set of questions for the This-or-That method.

Item 1	Which game did you like the most?
Item 2	Which game did you find a bit stupid?
Item 3	Which game would you like to get as a present?
Item 4	Which game did you find a bit boring?
Item 5	Which game would you like to play again?
Check	Sorry, I forgot. Which game did you like the most?

This method is comprised of 5 simple preference questions (see Table 1), upon which the preschooler chooses his or her preferred product among two alternatives (in our case APEX or DIESEL-X). The child can choose, simply by pointing, and thus without having to verbalize the answer. Children are also allowed to be indecisive in case they have no clear preference.

Normally, This-or-That includes a free play option (i.e. after the questions, the researcher tells the child that they have five minutes left, and that the child can spend the remaining time playing one out of the two alternatives), which is used as a validity check. In this case, there was no time for a free play option. Instead, we repeated the first question, namely "Which game do you like the most?"

A preference for DIESEL-X was scored as 1 for DIESEL-X and 0 for APEX, and vice versa (reversing item 2 and item 4). Hence, the maximum score a game can receive from one child is 5. A score of 0.5 for both games was given in case of indecisive answers. The last question in the set is similar to the first one, and was included the check the consistency in the child's answer. The reliability of the scale was assessed. Cronbach's alpha was .879 and could not be increased by dropping any of the items. Hence all items were retained. The correlation between the average score on all five items and the check item was found to be at .906 ($p < 0.01$).

Measurement of attention

Attention cannot reliably be measured post-hoc by asking preschoolers to fill out a self-report. It is recommended to base measurements of attention on behaviors that can be observed: behaviors such as activity and concentration on the task, or the absence of daydreaming, looking around, fidgeting, staring blankly [24].

Based on these recommendations, a scoring table for the level of attention was drafted (see Table 2). A score of 1 accords with no attention, a level of 5 accords with full attention. A score of 3 points to an alternation of periods of attention and inattention in equal amounts. Two researchers observed the child independently of each other, afterward the score was discussed and agreed upon. As the consensus was reached immediately, no intercoder reliability was computed.

Table 2. Scoring table for the level of attention

Scoring of attention	Description of behaviors that can be observed
1	The child has no attention for the task. The child performs no activity related to the task. The child is looking around, fidgeting or staring blankly. The child is fully distracted.
2	The child has almost no attention for the task. The child performs little activities related to the task. The child is mostly looking around, fidgeting, daydreaming or staring. Moments of concentration or activity are less than 20% of the time of the test period. Periods of distraction clearly dominate.
3	The child alternates periods of attention with periods without attention for the task, in more or less equal amounts. Moments of concentration and activity on the task alternate with moments of distraction.
4	The child has clear attention for the task. Only occasionally does the child show moments of distraction (less than 20%) by looking around, fidgeting, daydreaming or staring. Moments of concentration clearly dominate moments of distraction.
5	The child has full attention for the task, and is full y concentrated on the task. There are no moments of distraction where the child is looking around, fidgeting or staring blankly.

Performance on FM detection.
The procedure of the FM detection task was identical for APEX and DIESEL-X. A 2-up, 1-down, three-alternative forced-choice (3AFC) task was used to detect the threshold for frequency modulation. This threshold is defined as the minimum modulation depth a tone should have in order to distinguish it from non-modulated tones. The length of the stimuli was always 1000 ms, with a carrier frequency of 1 kHz and a modulation frequency of 2 Hz. The first trial started with a very large modulation depth (40 Hz) that was easily detectable. Step sizes between trials were initially large, and set to a minimum value after four reversals (see Figure 1). As the modulation depth decreases, it becomes harder and harder to distinguish the FM tone from the unmodulated tones. After eight reversals, the experiment ended, and the average value of the modulation depth at the last four reversals was taken as the detection threshold.

RESULTS

Game experience
The first hypothesis posited that DIESEL-X would provide a better game experience than APEX. Hence, preschoolers should show a preference for the DIESEL-X game.

This-or-That give an average score of 4.17 (SD = 1.51) for DIESEL-X on the maximum score of 5. Logically, the score for APEX was 0.83 (5 – 4.17). Ninety out of 95 preschoolers preferred DIESEL-X over APEX (i.e. they gave DIESEL-X a score of 3 on 5 or more). A one-sampled t-test confirmed that this preference was highly significant, $t(94) = 26.87$, $p < .001$. No order effect was found.

Measurement of attention
The second hypothesis posited that preschoolers would show greater attention when playing the DIESEL-X game, as compared to when playing the APEX game. Attention was scored based on a scoring table [24], see Table 2. Preschoolers that showed maximum attention received a score of 5, minimum attention was a score of 1.

For both games, attention was high. The average attention score for DIESEL-X was 4.69 (SD = .566), the attention score for APEX was 4.73 (SD = .591). However, we did not find a significant difference with respect to attention. No order effect was found.

Performance on FM detection
The third hypothesis posited that the DIESEL-X game would measure lower FM-thresholds than the APEX game. As aforementioned, the average value of the modulation depth at the last two reversals was taken as the detection threshold. The average modulation depth across all participants was 9.49 Hz (SD 5.08) for DIESEL-X, while it was 10.79 Hz (SD = 5.24) for APEX. Although this difference is relatively small, it is significant; $t(94) = 2.77$; ($p=0.007$). No interaction with the order in which the games were played was found.

DISCUSSION
Our results confirm the hypothesis that children preferred the game experience provided by the DIESEL-X game over the APEX game. Moreover, the results also confirm the hypothesis that the DIESEL-X game was able to measure smaller FM thresholds than the APEX game.

While the observed differences in thresholds may seem small, note that psychoacoustics thresholds reflect the capabilities of very basic processes in the ear and in the brain, and are considered as stable and hence non-trainable. This makes the results that can be obtained from the DIESEL-X game very different from games that aim at obtaining training or learning effects. If large differences would have been found, that would imply that the results of previous studies with APEX are invalid. The fact this study was actually able to obtain lower thresholds, is a significant result on its own, and a challenging finding for the domain of psychoacoustics.

Scientific merit of good game design
It is equally an important finding for the domain of game-based assessment. This study demonstrates that the quality of game design has a significant effect on the quality of the scientific measurement of psychometrics. Hence, serious games developed by researchers without formal training and insight in game design, risk to miss scientific effects.

We remark that research projects in the domain of serious games often stress the importance of the scientific quality of evidence-based studies. In such projects, there needs to be a balance between the resources spent on game design and development, and the resources spent on the clinical study to show effects. We like to stress that sufficient resources should still be reserved for serious game design, or the power that might be gained with a larger sample of respondents is wasted on poor design choices.

Unfortunately, our study does not reveal which aspect of DIESEL-X mainly contributes to the better performance, since many variables were changed between the APEX game and the DIESEL-X game. Future work should further concentrate on dissecting which game attributes are causal. However, given the fact that thresholds are considered stable, it might be hard to find 'significant' effects when isolating each design element by itself.

We also could not confirm our second hypothesis (i.e. preschoolers will pay greater attention to the DIESEL-X game than to the APEX game). Most likely, the scoring table was not the right measurement instrument for picking up differences in attention. There may have been a ceiling effect, as average scores of 4.69 (DIESEL-X), and 4.73 (APEX) were close to the maximum score of 5.

This ceiling effect may come across as surprising, given that in earlier studies, the researchers noted that it was hard to keep the child concentrated on the task with the APEX game [6, 16]. One possible explanation may be that in the current study only one characteristic (frequency modulation) was assessed for a duration of about 15 minutes, while typically in a classical experiment with APEX three or more assessments are carried out (e.g. amplitude modulation, gap

in noise detection, etc.), making the experiment last over 45 minutes. A second explanation is that researchers still actively encouraged children during the APEX test. They verbally and visually prompted the child, and they lauded the child in case of success. With respect to assessing attention, this may have skewed the results. In contrast, with DIESEL-X no researcher intervention was necessary during the entire test. The researchers in their report wrote: *"Because DIESEL-X was played without intervention of the researcher, and the child was playing with a headphone, as researchers we could only monitor the game play from a distance."* This quote does suggest that DIESEL-X enabled sustained attention without extra researcher intervention.

A caveat for game-based assessments

Our analysis also unearthed a fundamental problem that all game-based assessments need to overcome. Assessment games are measuring player characteristics or skills that are assumed to be stationary over the course of the game assessment. Hence, the core skill or player characteristic that the game assesses is by definition beyond the control of the player. This is potentially detrimental to feelings of learner control [18], i.e. the capacity of a player to influence or exert power over certain elements of the game. Learner control is essential to a sense of mastery or perceived competence but also for a sense of autonomy, both important ingredients of a game experience and of intrinsic motivation [15,23].

In the DIESEL-X game, to address the above problem, the choice was made to let players unlock new colors and gears. Players could then choose which color or gear pack the main character could take on in the remainder of the game. In this way, players were perhaps given a sense of control and freedom of choice. Observations confirmed that preschoolers particularly like this option. However, the suggestion that unlocking achievements and gear packs contributed to learner control, and hence positively influenced scientific measurements remains speculative. When questioned, preschoolers evaluate game experience in a one-dimensional way [53,55]. Consequently, in this study, we could nor isolate neither evaluate the separate effects of the different game attributes that might be responsible for the improvement in game experience and the resulting scientific measurement.

FUTURE WORK AND CONCLUSION

The presented work illustrates the feasibility of incorporating rather boring psychoacoustic assessments in a game environment that is highly valued by children. This game-based assessment also enabled to measure smaller FM thresholds. Moreover, our study demonstrates that not all games are equal when it comes down to the assessment of psycho-acoustic thresholds. The quality of game design has an effect on the quality of the scientific assessment of psychometrics. Serious games developed by researchers without formal training and insight in game design, risk to miss scientific effects. Better game experiences lead to longer sustained attention, in turn leading to more accurate measurements of the test participant's capabilities, in sum resulting in a better screening tool.

In the near future, further longitudinal studies are planned in which the performance on DIESEL-X is compared between children diagnosed with dyslexia (which cannot happen before the age of 8 years) and children with normal reading and writing skills. The ultimate goal is to obtain a tool that can detect a high risk for developing dyslexia, at preschool age. Therefore, the game will be tested with children at different ages, including dyslectic children, in order to find predictive measures that can be obtained by playing the game.

ACKNOWLEDGEMENTS

We like to thank all the preschoolers and schools that participated in the study. In addition we like to thank game designers and developers that contributed to DIESEL-X, namely Jelle Husson and Lieven Van den Audenaeren. This paper is based on work supported by IWT, the government agency for Innovation by Science and Technology under grant 100767.

REFERENCES

1. Peter J Bailey and Margaret J Snowling. 2002. Auditory processing and the development of language and literacy. *Br Med Bull* 63, 1, 135–146. http://doi.org/10.1093/bmb/63.1.135
2. Albert Bandura and Edwin A. Locke. 2003. Negative self-efficacy and goal effects revisited. *Journal of applied psychology* 88, 1, 87.
3. Roy F. Baumeister, Ellen Bratslavsky, Catrin Finkenauer, and Kathleen D. Vohs. 2001. Bad is stronger than good. *Review of general psychology* 5, 4, 323.
4. Wendy L. Bedwell, Davin Pavlas, Kyle Heyne, Elizabeth H. Lazzara, and Eduardo Salas. 2012. Toward a Taxonomy Linking Game Attributes to Learning: An Empirical Study. *Simulation & Gaming*, 1046878112439444. http://doi.org/10.1177/1046878112439444
5. Bart Boets, Pol Ghesquière, Astrid van Wieringen, and Jan Wouters. 2007. Speech perception in preschoolers at family risk for dyslexia: Relations with low-level auditory processing and phonological ability. *Brain and Language* 101, 1, 19–30. http://doi.org/10.1016/j.bandl.2006.06.009
6. Bart Boets, Jan Wouters, Astrid van Wieringen, B Desmedt, and Pol Ghesquière. 2008. Modelling relations between sensory processing, speech perception, orthographic and phonological ability, and literacy achievement. *Brain and Language* 106, 1, 29–40. http://doi.org/10.1016/j.bandl.2007.12.004
7. Bart Boets, Jan Wouters, Astrid van Wieringen, and Pol Ghesquière. 2006. Auditory temporal information processing in preschool children at family risk for dyslexia: Relations with phonological abilities and

developing literacy skills. *Brain and Language* 97, 1, 64–79. http://doi.org/10.1016/j.bandl.2005.07.026
8. Bart Boets, Jan Wouters, Astrid van Wieringen, and Pol Ghesquière. 2006. Coherent motion detection in preschool children at family risk for dyslexia. *Vision Research* 46, 4, 527–535. http://doi.org/10.1016/j.visres.2005.08.023
9. Bart Boets. 2006. Early literacy development in children at risk for dyslexia. A longitudinal study of the general magnocellular theory.
10. L. Bradley and P. E. Bryant. 1983. Categorizing sounds and learning to read: a causal connection. *Nature* 301, 5899, 419–421. http://doi.org/10.1038/301419a0
11. Amy Bruckman. 1999. Can educational be fun? *Game Developers Conference '99*.
12. Véronique Celis, Jelle Husson, Vero Vanden Abeele, et al. 2013. Translating preschoolers' game experiences into design guidelines via a laddering study. *Proceedings of the 12th International Conference on Interaction Design and Children*, ACM, 147–156. http://doi.org/10.1145/2485760.2485772
13. Diana I. Cordova and Mark R. Lepper. 1996. Intrinsic motivation and the process of learning: Beneficial effects of contextualization, personalization, and choice. *Journal of educational psychology* 88, 4, 715.
14. Mihaly Csikszentmihalyi. 1990. *Flow: The Psychology of Optimal Experience*. Harper and Row, New York.
15. E. L. Deci and R. M. Ryan. 1985. *Intrinsic motivation and self-determination in human behavior*. Springer.
16. M. E. Farmer and R. M. Klein. 1995. The evidence for a temporal processing deficit linked to dyslexia: A Review. *Psychonomic bulletin & review* 2, 4, 460–493.
17. Angela Fawcett and Rod Nicolson. 1995. *Dyslexia in Children: Multidisciplinary Perspectives*. Harvester Wheatsheaf.
18. Rosemary Garris, Robert Ahlers, and James E. Driskell. 2002. Games, Motivation, and Learning: A Research and Practice Model. *Simulation & Gaming* 33, 4, 441–467. http://doi.org/10.1177/1046878102238607
19. D. C. M. Gersons-Wolfensberger and Wied A. J. J. M. Ruijssenaars. 1997. Definition and Treatment of Dyslexia: A Report by the Committee on Dyslexia of the Health Council of the Netherlands. *Journal of Learning Disabilities* 30, 2, 209–13.
20. Luc Geurts, Vero Vanden Abeele, Véronique Celis, et al. 2015. DIESEL-X: A game-based tool for early risk detection of dyslexia in preschoolers. In *Describing and Studying Domain-Specific Serious Games*, Joke Torbeyns, Erno Lehtinen and Jan Elen (eds.). Springer.
21. M. P. Jacob Habgood and Shaaron E. Ainsworth. 2011. Motivating Children to Learn Effectively: Exploring the Value of Intrinsic Integration in Educational Games. *Journal of the Learning Sciences* 20, 2, 169–206. http://doi.org/10.1080/10508406.2010.508029
22. Sini Hintikka, Aro Mikko, and Lyytinen, Heikki. 2005. Computerized training of the correspondences between phonological and orthographic units. *Written Language and Literacy* 8, 2, 79–102.
23. Wijnand IJsselsteijn, Yvonne A.W. de Kort, Karolien Poels, Audrius Jurgelionis, and Victoria Bellotti. 2007. Characterising and Measuring User Experiences in Digital Games. ACM Press.
24. Peter F. de Jong. 1995. Assessment of attention: Further validation of the star counting test. *European Journal of Psychological Assessment* 11, 2, 89–97. http://doi.org/10.1027/1015-5759.11.2.89
25. Johan Laneau, Bart Boets, Marc Moonen, Astrid van Wieringen, and Jan Wouters. 2005. A flexible auditory research platform using acoustic or electric stimuli for adults and young children. *Journal of Neuroscience Methods* 142, 1, 131–136. http://doi.org/10.1016/j.jneumeth.2004.08.015
26. Nicole Lazarro. 2004. Why We Play Games: Four Keys to More Emotion Without Story. XEO design. Retrieved March 1, 2009 from http://www.xeodesign.com/whyweplaygames/xeodesign_whyweplaygames.pdf
27. Dianne L. Lefly and Bruce F. Pennington. 1991. Spelling Errors and Reading Fluency in Compensated Adult Dyslexics. *Annals of Dyslexia* 41, 143–62.
28. H. Levitt. 1971. Transformed up-down methods in psychoacoustics. *The Journal of the Acoustical society of America* 49, 2B, 467–477.
29. A. Locke and G. Latham. 1989. *A Theory of Goal Setting and Task Performance*. Prentice-Hall.
30. Heikki Lyytinen and Jane Erskine. 2006. Early Identification and Prevention of Reading Problems. *Encyclopedia on Early Childhood Development*. Retrieved from http://www.enfant-encyclopedie.com/Pages/PDF/Lyytinen-ErskineANGxp.pdf
31. Thomas W. Malone and M. R. Lepper. 1987. Making learning fun: A taxonomy of intrinsic motivations for learning. *Aptitude, learning, and instruction* 3, 223–253.
32. G. M. McArthur and D. V. Bishop. 2001. Auditory perceptual processing in people with reading and oral language impairments: current issues and recommendations. *Dyslexia (Chichester, England)* 7, 3, 150–170. http://doi.org/10.1002/dys.200
33. E. McAuley, T. Duncan, and V. V. Tammen. 1989. Psychometric properties of the Intrinsic Motivation Inventory in a competitive sport setting: a confirmatory factor analysis. *Research quarterly for exercise and sport* 60, 1, 48.
34. Peter Menell, Ken I. McAnally, and John Stein. 1999. Psychophysical Sensitivity and Physiological Response to Amplitude Modulation in Adult Dyslexic Listeners. *J Speech Lang Hear Res* 42, 4, 797–803.
35. Dario Di Nocera, Alberto Finzi, Silvia Rossi, and Mariacarla Staffa. 2014. The role of intrinsic motivations in attention allocation and shifting.

Frontiers in Psychology 5. http://doi.org/10.3389/fpsyg.2014.00273
36. Louise E. Parker and Mark R. Lepper. 1992. Effects of fantasy contexts on children's learning and motivation: Making learning more fun. *Journal of Personality and Social Psychology* 62, 4, 625.
37. Michael I. Posner. 1980. Orienting of attention. *Quarterly journal of experimental psychology* 32, 1, 3–25.
38. J. C. Read, S. J. MacFarlane, and Chris Casey. 2002. Endurability, engagement and expectations: Measuring children's fun. *Interaction design and children*, Shaker Publishing Eindhoven, 1–23. Retrieved April 5, 2015 from http://chici.org/references/endurability_engagement.pdf
39. Janet C. Read. 2008. Validating the Fun Toolkit: An Instrument for Measuring Children's Opinions of Technology. *Cogn. Technol. Work* 10, 2, 119–128. http://doi.org/10.1007/s10111-007-0069-9
40. Katrina E. Ricci, Eduardo Salas, and Janis A. Cannon-Bowers. 1996. Do computer-based games facilitate knowledge acquisition and retention? *Military Psychology* 8, 4, 295–307. http://doi.org/10.1207/s15327876mp0804_3
41. Lucy J. Robinson, Lucy H. Stevens, Christopher J. D. Threapleton, Jurgita Vainiute, R. Hamish McAllister-Williams, and Peter Gallagher. 2012. Effects of intrinsic and extrinsic motivation on attention and memory. *Acta Psychologica* 141, 2, 243–249. http://doi.org/10.1016/j.actpsy.2012.05.012
42. Paul Rozin and Edward B. Royzman. 2001. Negativity Bias, Negativity Dominance, and Contagion. *Personality and Social Psychology Review* 5, 4, 296–320. http://doi.org/10.1207/S15327957PSPR0504_2
43. Sally E. Shaywitz. 1998. Dyslexia. *New England Journal of Medicine* 338, 5, 307–312. http://doi.org/10.1056/NEJM199801293380507
44. Gavin Sim and Matthew Horton. 2012. Investigating Children's Opinions of Games: Fun Toolkit vs. This or That. *Proceedings of the 11th International Conference on Interaction Design and Children*, ACM, 70–77. http://doi.org/10.1145/2307096.2307105
45. Margaret J. Snowling. 2000. *Dyslexia*. Wiley Blackwell.
46. J.B. Talcott, C. Witton, M F McLean, et al. 2000. Dynamic sensory sensitivity and children's word decoding skills. *Proceedings of the National Academy of Sciences of the United States of America* 97, 6, 2952–2957. http://doi.org/10.1073/pnas.040546597
47. J.B. Talcott and C. Witton. 2002. A sensory linguistic approach to the development of normal and impaired reading skills. In *Neuropsychology and cognition series. Basic functions of language and language disorders*. Dordrecht, Netherlands: Kluwer Academic Publishers.
48. Ed S. Tan and Jeroen Jansz. 2007. The Game Experience. *Product Experience*.
49. Vero Vanden Abeele, Bob De Schutter, Luc Geurts, et al. 2012. P-III: A Player-Centered, Iterative, Interdisciplinary and Integrated Framework for Serious Game Design and Development. In *Serious Games: The Challenge*, Stefan Wannemacker, Sylke Vandercruysse and Geraldine Clarebout (eds.). Springer Berlin Heidelberg, 82–86. Retrieved October 9, 2012 from http://www.springerlink.com/content/k66727215741j3t0/abstract/
50. C. Witton, J.B. Talcott, P C Hansen, et al. 1998. Sensitivity to dynamic auditory and visual stimuli predicts nonword reading ability in both dyslexic and normal readers. *Current Biology: CB* 8, 14, 791–797.
51. Wolfshead. 2007. Bushnell's Theorem: Easy to Learn, Difficult to Master | Wolfshead Online. Retrieved April 8, 2010 from http://www.wolfsheadonline.com/?p=81#291dc
52. Bieke Zaman, Vero Vanden Abeele, and Dirk De Grooff. 2013. Measuring product liking in preschool children: An evaluation of the Smileyometer and This or That methods. *International Journal of Child-Computer Interaction* 1, 2, 61–70. http://doi.org/10.1016/j.ijcci.2012.12.001
53. Bieke Zaman, Vero Vanden Abeele, and Dirk De Grooff. 2013. Measuring product liking in preschool children: An evaluation of the Smileyometer and This or That methods. *International Journal of Child-Computer Interaction* 1, 2, 61–70. http://doi.org/10.1016/j.ijcci.2012.12.001
54. Bieke Zaman and Vero Vanden Abeele. 2007. How to Measure the Likeability of Tangible Interaction with Preschoolers. *Proceedings of CHI.NL*, Infotec Nederland BV Woerden, 57–59. Retrieved March 20, 2009 from http://soc.kuleuven.be/com/mediac/cuo/admin/upload/Zaman_Abeele_chiformat.pdf
55. Bieke Zaman. 2009. Introducing a Pairwise Comparison Scale for UX Evaluations with Preschoolers. In *Human-Computer Interaction – INTERACT 2009*. 634–637. Retrieved October 6, 2009 from http://dx.doi.org/10.1007/978-3-642-03658-3_68